DOCUMENT RESUME

ED 363 272 IR 016 259

AUTHOR Parshall, Cynthia G.; Kromrey, Jeffrey D.
TITLE Computer Testing versus Paper-and-Pencil Testing: An Analysis of Examinee Characteristics Associated with Mode Effect.
PUB DATE Apr 93
NOTE 41p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Age Differences; College Entrance Examinations; College Students; Comparative Testing; *Computer Assisted Testing; Computer Literacy; Demography; Ethnic Groups; Higher Education; *Individual Characteristics; Intermode Differences; Pilot Projects; Racial Differences; Sex Differences; *Student Experience; *Test Format; Test Wiseness
IDENTIFIERS Graduate Record Examinations; *Paper and Pencil Tests; *Testing Effects

ABSTRACT

This paper studies whether examinee characteristics are systematically related to mode effect across paper and computer versions of the same instrument, using data from the Graduate Record Examination (GRE) of the Educational Testing Service in its Computer-Based Testing Pilot Study of 1991. The following characteristics of 1,114 examinees were studied as contributors to mode effect: (1) demographic variables (gender, racial/ethnic background, and age); (2) computer use variables (variety and frequency of computer experience, frequency of mouse use, and test mode preference); and (3) test-taking strategy variables (strategy preference, and tendency to omit or review items). The typical method found in comparability literature was used, treating performance across paper and computer versions as a continuous dependent variable. Because a mode effect in a small subset of examinees could be masked, a method that isolated examinees most affected by test mode was used. For this method, mode effect was treated as a three-level, categorical, independent variable. Data demonstrate mode effect and support the conception of a small subset of examinees whose performance was more affected by mode than that of the total sample. The search for examinee characteristics that explain occurrence of mode effect, however, yielded inconsistent results, with only weak relationships to mode effect. Three figures and 21 tables present study findings. (Contains 31 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 363 272

Computer Testing versus Paper-and-Pencil Testing:

An Analysis of Examinee Characteristics Associated with Mode Effect

Cynthia G. Parshall

American College Testing

Jeffrey D. Kromrey

Department of Educational Measurement and Research

University of South Florida

Paper presented at the annual meeting of the American Educational Research Association, April 12-16, 1993, Atlanta, GA.

TM019259

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Cynthia G. Parshall

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Abstract

A variety of computer test applications have been developed in recent years and computer-based testing has proved feasible in terms of the potential for adequate degrees of reliability and validity. A more difficult proposition is the direct, concurrent equivalence of computer and paper versions of a given examination. Many instances have been noted of score differences across the two test modes, commonly termed mode effect. The purpose of this research was to determine whether examinee characteristics are systematically related to mode effect across paper and computer versions of the same examinations.

The data for this study were obtained from the Educational Testing Service's Graduate Record Examination (GRE) Program. The GRE Program conducted a Computer-Based Testing Pilot study in the Fall of 1991; secondary analyses of the results of that pilot study were conducted for purposes of this research.

The examinee characteristics that were investigated as possible contributors to mode effect included: (1) demographic variables (gender, racial/ethnic background, and age), (2) computer use variables (variety of computer experience, frequency of computer use, frequency of mouse use, test mode preference), and (3) test taking strategy variables (test strategy preference, tendency to omit items, tendency to review items).

The data were analyzed through two methods. First, the typical method found in the literature on comparability studies was employed. In this method, performance across the paper and computer version is treated as a continuous, dependent variable. Secondly, because a mode effect in a small subset of examinees could be masked when all subjects are considered together, a method which isolated those examinees most affected by test mode was also used. For this method, mode effect was treated as a three-level, categorical, independent variable.

The data in this study demonstrated mode effect and supported the conception of a small subset of examinees whose performance was more affected by test mode than was that of the total sample of examinees. The search for examinee characteristics that explain the occurrence of mode effect, however, yielded inconsistent results. The variables investigated in this study showed only relatively weak relationships to mode effect.

Introduction

When computer-based tests are developed from pre-existing paper and pencil tests information is usually required concerning the comparability of examination scores obtained in one mode to scores obtained in the other mode. Bunderson, Inouye, & Olsen, (1989) state that, for the field of computer-based testing, "The fundamental research question...is the equivalence of scores between a computerized version of a test and the original version" (p. 378).

Comparability of scores obtained on computer and paper versions of an examination is needed to establish validity across forms and to ensure that norms which were developed from a paper-administered version are useable for the computer version. The comparability of test scores is also a vital issue because of the need to provide fair testing for all examinees. According to the Joint Committee's Standards for Educational and Psychological Testing, when alternate forms of an examination are in use, "it should be a matter of indifference to anyone taking the test or to anyone using the test results" which form of the examination was administered (Committee to Develop Standards for Educational and Psychological Testing [Joint Committee], 1985, p. 31). When test mode affects performance, this is clearly not the case.

Mazzeo and Harvey (1988) conducted an extensive review of comparability studies, summarizing research on computerized personality, ability, and achievement tests. After finding that mean score differences across mode frequently occurred, their conclusion was that "a critical issue...is the equivalence of computer-generated scores to paper-and-pencil scores" (p. 2).

Although comparability can be approached from the total test score perspective, or from the item level perspective, this study analyzed comparability from the examinee mode effect perspective. Some researchers suggest that small overall differences in performance across test modes may actually be the result of more sizable differences affecting only a small number of examinees (Wise, Barnes, Harvey, & Plake, 1989; Wise & Plake, 1989). This comparability study was designed to look at subgroups of examinees in order to isolate the scores of those who appear to be most affected by test mode, and then to attempt to find the examinee characteristics which distinguish those examinees who are affected by test mode from those who are not.

Examinee Characteristics as Contributors to Mode Effect

Research on the relationship between various examinee characteristics and mode effect has not been conclusive. Gender, race or ethnic background, and age are among the demographic attributes which have been investigated (Buhr & Legg, 1989; Johnson & Mihal, 1973; Johnson & White, 1980; Llabre & Froman, 1987; Moe & Johnson, 1988; Sorensen, 1985). Although decisive evidence of the relationship between these characteristics and mode effect has not been obtained, grounds for concern can be found in the results of national surveys on equity of computer access (Becker & Sterling, 1987; McPhail, 1985). Across elementary and secondary levels, schools with predominantly minority students were more likely to have no computers or to have higher student-per-computer ratios. Within schools, males more than females tended to take part in optional computer courses and computer activities. To the extent that certain subgroups have unequal access to computers, concerns about the consequences of amount of computer experience on performance under computer-administered test conditions become applicable.

Several aspects of examinee interactions with computers have been investigated as potential sources of mode effect; thus far, clear-cut results have not been obtained (Eaves & Smith, 1986; Lee, 1986). A pattern of lower scores for examinees with less computer experience is frequently seen, although the score differences are often not statistically significant (Wise et al., 1989). Several studies have investigated the relationship between examinee performance on computer tests and a stated preference for paper or computer versions of a test, but no direct relationship has been consistently found (Bugbee & Bernt, 1990; Buhr & Legg, 1989; Koch & Patience, 1978).

Another area of research is that of the relationship between test taking strategy preferences and computer test scores (Rocklin & O'Donnell, 1987; Wise & Plake, 1989). The operation of taking a test on computer is different in certain procedural ways; whether all examinees are able to adapt to those differences without impact on their performance is not certain. Specifically, the interaction of examinees' test-taking strategies with test flexibility (i.e., the presence or absence of features which enable a computer test-taker to omit and revise answers) appears to be important (Green, 1991; Spray, Ackerman, Reckase, & Carlson, 1989; Sachar & Fletcher, 1978; Ward, Hooper, & Hannafin, 1989). Wise and Plake (1990) have stated

> ...individual examinees vary greatly in their item skipping, answer reviewing, and answer changing behavior; yet to be ascertained is whether the scores of examinees who strongly prefer to return to items would be affected by denying them that opportunity (p. 8).

Methods

This study was a secondary analysis of data from the Educational Testing Service's (ETS) pilot study of a computer version of the Graduate Record Examination (GRE) General Test. During the standard, paper administration of the GRE conducted in October, 1991, examinees who took a specific test form were given the opportunity to enlist in the pilot study. Examinees were offered a $50 honorarium for participating and were given the option, if the pilot study proved successful, to have their scores from the computer administration of the examination added to their score records. Since subjects were thus offered an opportunity to better their scores in a high-stakes test environment, motivated participation was encouraged.

Examinees from the October administration who agreed to participate were then re-tested on an alternate test form during November or December of 1991. A small number of examinees were given this second test in paper form, to provide ETS with cross-checks in the research design. For the majority of examinees, however, the second test was administered on computer. Only those examinees who took both a paper and a computer version are considered in this secondary analysis. The total number of examinees whose scores were used in the secondary analysis was 1,114. This sample of examinees is typical of GRE test takers in terms of test scores, gender, and ethnicity (ETS, 1992).

Instruments

The standard, paper-version of the GRE General Test consists of two sections each of Quantitative, Verbal, and Analytical scales. One non-operational section is typically also administered in order to pre-test new items; scores on this seventh section were not used in this study. KR-20 estimates of reliability for the test form used in the pilot study yielded a .91 estimate of reliability for the Verbal scale, .92 for the Quantitative scale, and .88 for the Analytical scale (ETS, 1992).

The computer test was completely parallel to the paper test in terms of standard sections, item types, and numbers of items. The seventh section in this version was a non-flexible Verbal section and will be discussed in further detail later. Because the computer took two to three seconds to access and present each

item, the time allotment for each section on the computer version was 32 minutes, rather than 30 minutes. The computer test was administered on IBM PCs with VGA monitors and ran under Microsoft Windows. Administration of the computer test was immediately preceded by a computer tutorial on the use of the test administration software. KR-20 estimates on this test form resulted in .92 for both the Verbal and Quantitative scales and .89 for the Analytical scale. (ETS, 1992).

Demographic information on examinees was obtained through a questionnaire completed by examinees as part of the registration process. Information about examinees' prior computer experience and their reactions to the computer version were collected through a computer test survey, administered to examinees after they had completed the computer test.

Residual Difference Scores and Methods of Analysis

The presence of a mode effect was determined by evaluating the difference in examinee performance across computer and paper versions of the parallel test forms. For the most part, these differences were examined in terms of residual scores rather than difference scores, because of the tendency of simple difference scores to be unreliable (Linn, 1988). The residual difference scores were calculated as the difference between the examinees' predicted computer test scores (based on a linear regression of the computer version scores on the paper version scores) and their actual computer test scores. Thus, a negative residual difference score would suggest that an examinee performed less well on the computer version than his or her own paper version score would predict. The residual scores were computed separately for each test scale.

Two methods were used in this study to investigate the occurrence of test mode effect. The first method (Method 1) followed the more typical process of assessing the presence of mode effect through comparisons of examinee performance across the two modes (computer and paper versions). For this method examinees' residualized difference scores on the Verbal, Quantitative, and Analytical scales were used as continuous, dependent variables.

An alternative method was also used (Method 2), in order to investigate the hypothesis that mode effect occurs predominantly in a small subset of examinees (Wise et al., 1989). For this method, the approach was to isolate the scores of those subjects whose performance appears to be most affected by mode of test

administration. Based on their residual difference scores, subjects were therefore classified according to three levels of mode effect (mode effect favoring the paper test score, no mode effect, and mode effect favoring the computer test score) using plus or minus one standard deviation as the criteria. Examinees whose residual scores were within one standard deviation of zero were classified as having no mode effect; those with residuals one standard deviation above the mean were assigned to the computer mode effect group; and those with residual difference scores one standard deviation below the mean were assigned to the paper mode effect group. One standard deviation was set as the cut-off in order to provide for sufficient sample size in the two extreme mode effect groups; since the residual difference scores were approximately normally distributed, roughly two-thirds of the sample were assigned to the no mode effect group. These mode effect groups were obtained separately for the Verbal, Quantitative, and Analytical scales.

It is important to realize that the fact that extreme residuals were found under Method 2 is not at issue; the presence of extreme values is a natural occurrence when residual scores are computed. Instead, the point of interest under Method 2 is the relationship between the mode effect groupings and examinee characteristics.

Results

Omnibus Test for Mode Effect

The first analysis conducted was to determine whether an overall mode effect was present in the data. Hotelling's T^2 was performed on the difference between scale means across test mode. Simple difference scores (computer version mean score minus paper version mean score) on the three scales were used as the dependent measures in this analysis. A significant result was found (T^2 = 1,728.36, $p < .01$). Follow-up univariate t-tests on each examination scale were also found significant: Verbal, t (1112) = -2.22, $p < .05$; Quantitative, t (1112) = -40.40, $p < .01$; and Analytical, t (1112) = -11.76, $p < .01$.

An examination of test means (Table 1) indicates that, across all three test scales, examinees performed significantly better on the computer version of the examination than on the paper version of the same test. However, the fact that test mode and test order were perfectly confounded in the original design of the computer test pilot study makes interpretation of these omnibus results difficult. The overall improvement in scores on the

computer test could have resulted from practice effects rather than test mode differences. This is the conclusion which ETS has made about the test, based in part on additional data collected during the pilot study which were not used in this secondary analysis (ETS, 1992). However, even if retest effects caused the overall improvement in scores on the computer version of the examination, the large residual difference scores obtained by some examinees is not addressed. Spray et al. (1989) have suggested that item level differences across test mode may cancel each other out on a total test score; in the same way, positive and negative differences in examinee performance across test mode may cancel each other out at the total group level. The differences between mean scores across levels of the mode effect group variable (Table 2) suggest that on each examination scale, a subset of examinees performed very differently across test mode. The remainder of analyses conducted were aimed at determining examinee characteristics which define those examinees.

Table 1

Means and Standard Deviations of Test Scores by Scale and Test Mode

Scale	No. of Items	Test Mode: Paper	Test Mode: Computer
Verbal	76		
M		49.93	50.31
SD		12.42	12.05
Quantitative	60		
M		36.44	42.02
SD		10.81	10.59
Analytical	50		
M		32.21	33.96
SD		8.49	8.40

Note. N = 1114.

Table 2

Mean Scores and Standard Deviations by Test Mode, Mode Effect Group, and Scale

Test Mode	Mode Effect Group: Computer	No Effect	Paper
	Verbal (n of items = 76)		
	n = 159	n = 788	n = 167
Computer			
M	56.50	51.31	39.71
SD	9.53	11.62	9.62
Paper			
M	47.92	50.87	47.39
SD	11.64	12.89	10.17
	Quantitative (n of items = 60)		
	n = 105	n = 907	n = 101
Computer			
M	46.55	42.82	30.30
SD	6.89	10.27	8.77
Paper			
M	32.55	37.35	32.36
SD	7.88	11.09	8.99
	Analytical (n of items = 50)		
	n = 94	n = 915	n = 105
Computer			
M	39.61	34.63	23.04
SD	5.55	7.5	8.41
Paper			
M	29.03	32.83	29.62
SD	6.91	8.40	9.54

Demographic Variables

The next analysis conducted was an extensive examination of the relationship between mode effect and examinee demographic characteristics of gender, racial/ethnic background, and age. Sixty-one percent of the examinees were female, and 39% were male. The variable age was a dichotomy, with 18% of the examinees categorized as "Older" (those 30 years of age and older) and 82% categorized as "Younger". Table 3 provides the frequencies of examinees by race.

Table 3

Frequency Distribution of Examinee Sample by Race

Race	Frequency	Percent
Asians	74	6.8
Blacks	87	8.0
Hispanics	50	4.6
Whites	849	77.6
Other	34	3.1

Note. N = 1094.

Method 1. To determine if there were test mode effects due to examinee characteristics, data were subjected to a three-factor MANOVA (Table 4). Significant main effects were found for gender, and for race; significant interaction effects were found for gender by race, and for race by age. The three-way gender by race by age interaction was not significant.

Table 4

MANOVA of Residual Difference Scores on Verbal, Quantitative, and Analytical Scales by Gender, Race, and Age

Source	df	Wilks' Lambda	F
Gender	3, 1067	.99	4.33*
Race	12, 2823	.97	2.92*
Age	3, 1067	.99	2.45
Gender * Race	12, 2823	.97	2.57*
Gender * Age	3, 1067	1.00	1.24
Race * Age	12, 2823	.97	2.95*
Gender * Race * Age	12, 2823	.99	1.24

* $p < .01$.

Follow-up three-factor ANOVAs were conducted for each of the three test scales. Table 5 presents the ANOVA on the Verbal residual difference scores. Significant main effects were found for gender and race; significant interactions were found for gender by race, and race by age. Dunn's multiple comparison test was used as a follow-up test to each of the significant interaction effects, as relatively few pairwise comparisons were planned.

A visual examination of Figure 1 (a plot of the interaction effect between race and gender) suggests that at some levels of race, males' residual difference scores on the Verbal scale were negatively impacted compared to females. However, no individual pairwise comparisons were found to be significant for the gender by race interactions. The Asian, Black, and Hispanic levels of race were collapsed and categorized as "non-White" in order to analyze the effect across that group. (The category "Other" was disregarded at this point, since additional data about the ethnic make-up of examinees in that level were unavailable). When the comparison for the White versus non-White level of race was conducted, a significant difference was found between mean scores for males and females (Dunn's = 3.59, $p < .01$). This suggests that non-White males' performance on the computer version of the test was less than their performance on the paper version predicted, as compared to non-White females. The

significant main effects for gender and race could not be interpreted directly because of the significant interaction effects.

Table 5

Analysis of Variance of Residual Difference Scores by Gender, Race, and Age on Verbal Scale

Source	df	MS	F
Gender (G)	1	135.24	4.99*
Race (R)	4	113.57	4.19**
Age (A)	1	102.96	3.80
GR	4	85.42	3.15*
GA	1	61.38	2.26
RA	4	152.69	5.63**
GRA	4	54.95	2.03
Error (S/GRA)	1069	27.12	
Total	1088		

* $p < .05$. ** $p < .01$.

A plot of the interaction effect between race and age is shown in Figure 2. A significant difference was found between means for older and younger Blacks (Dunn's = 3.50, $p < .01$). This suggests that older Blacks performed less well on the computer version of the examination than their paper version scores predicted, as compared to younger Blacks. (Very low n's in the Asian, Hispanic, and Other levels of race by age may have reduced power for these post-hoc analyses.)

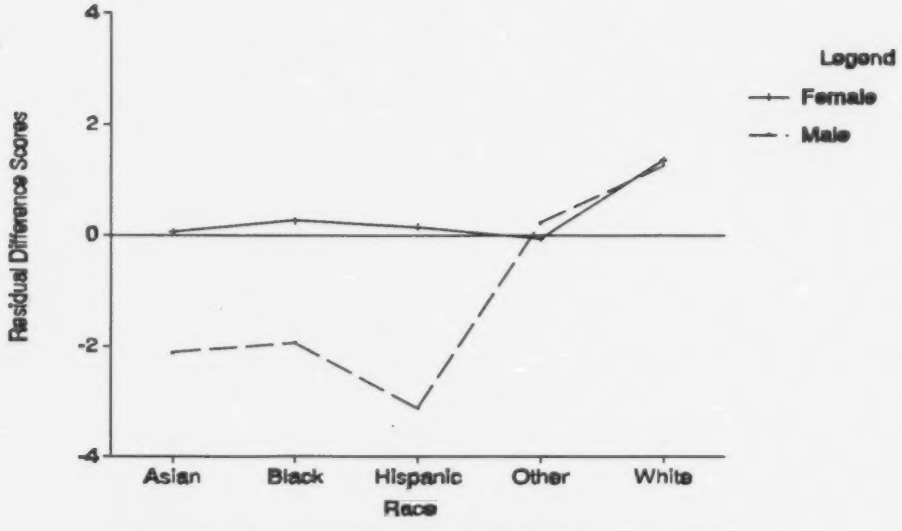

Figure 1: Interaction of Race and Gender on Verbal Residual Difference Scores.

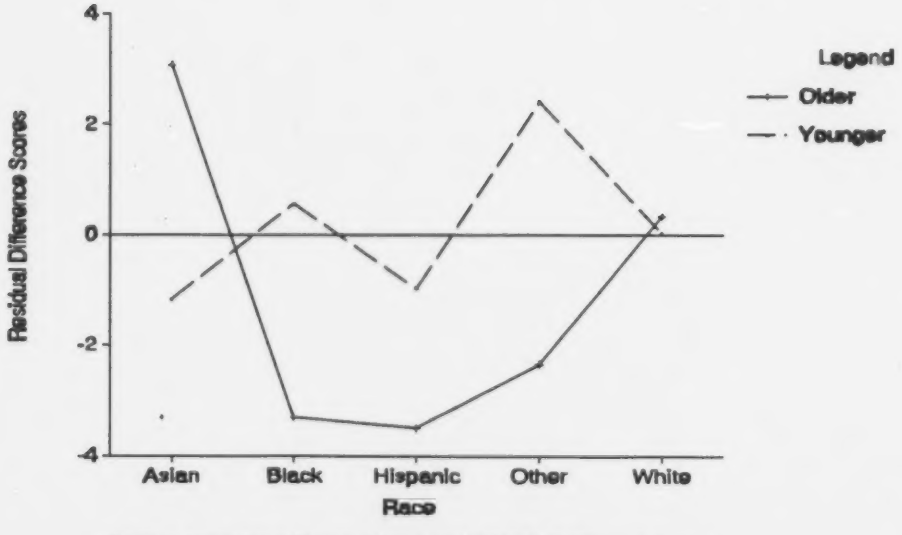

Figure 2: Interaction of Race and Age on Verbal Residual Difference Scores.

The three factor ANOVA on the Quantitative residual difference scores (Table 6) revealed a significant main effect for gender and a significant interaction effect between race and age. A plot of this interaction effect is shown in Figure 3. Again, the only pairwise comparison that was significant at the .05 hypothesiswise level was for Blacks (Dunn's = 3.34, $p < .01$). The same pattern held here as in the Verbal scale: older Blacks had a significantly lower mean residual difference score than did younger Blacks. Relative to the significant main effect for gender, males had a significantly higher mean residual difference score on the Quantitative scale than did females. This would suggest that, on the Quantitative scale, the overall performance of males on the computer version was higher than would be expected based on their paper version scores. For females, the average performance on the computer version was less than would be expected, based on their own paper version performance.

Table 6

Analysis of Variance of Residual Difference Scores by Gender, Race, and Age on Quantitative Scale

Source	df	MS	F
Gender (G)	1	128.05	6.70**
Race (R)	4	28.13	1.47
Age (A)	1	31.26	1.64
GR	4	37.80	1.98
GA	1	9.03	0.47
RA	4	51.15	2.68*
GRA	4	12.50	0.65
Error (S/GRA)	1069	19.11	
Total	1088		

* $p < .05$. ** $p < .01$.

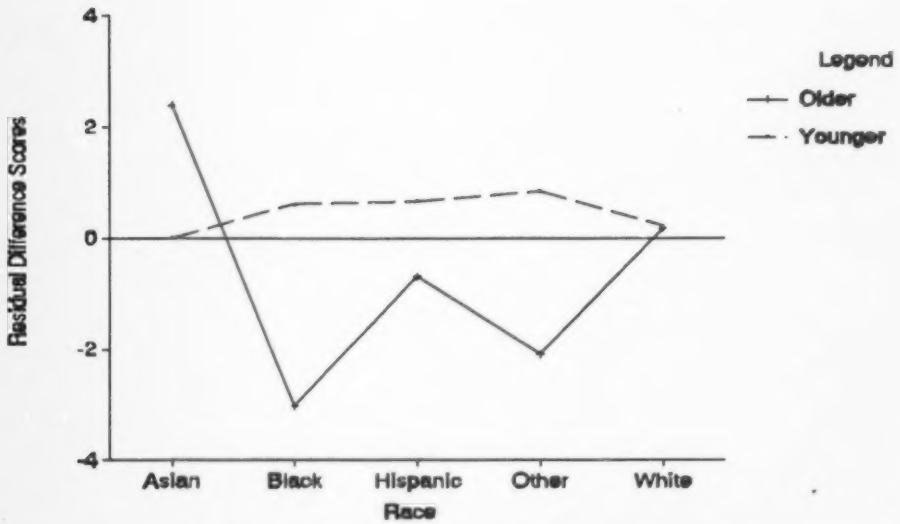

Figure 3: Interaction of Race and Age on Quantitative Residual Difference Scores.

The three factor ANOVA of gender, racial/ethnic background, and age on Analytical residual difference scores indicated a significant main effect for race (Table 7); no other significant effects were observed. Since all possible pairwise comparisons of the race variable were to be performed, Tukey's multiple comparison test was selected. Tukey's post hoc test on the main effect for race revealed a significant difference between Blacks and Asians ($p < .05$). Asians had a significantly greater mean residual score in the Analytical scale than did Blacks, indicating that Asians performed better on the computer version than their paper version scores would predict, as compared to Blacks.

Table 7

Analysis of Variance of Residual Difference Scores by Gender, Race, and Age on Analytical Scale

Source	df	MS	F
Gender (G)	1	0.19	0.01
Race (R)	4	94.03	4.46*
Age (A)	1	71.69	3.40
GR	4	46.48	2.21
GA	1	19.66	0.93
RA	4	42.90	2.04
GRA	4	17.76	0.84
Error (S/GRA)	1069	21.06	
Total	1088		

* $p < .01$.

Method 2. Under the Method 2 approach in this study, residual difference scores were used to categorize examinees according to level of mode effect. In order to investigate the possibility that mode effect groups were defined by crossed demographic variables (e.g. younger Hispanic females), log-linear analyses were conducted. The log-linear analyses performed were asymmetrical in nature; that is to say, only those associations affecting the mode effect group were of interest. The demographic variables of gender, racial/ethnic background, and age group served as the explanatory variables in these analyses, while mode effect group was the categorical respondent variable.

In log-linear analysis, the initial analysis provides a screening of four families of models. First, the null model (Model 0) is tested, in which the marginal frequencies of each variable are used to establish expected frequencies in each cell. Second, the main effects only model is tested (Model 1), in which the independent effects of the three explanatory variables contribute to the expected frequencies. Third, the main effects and two-way interaction model is tested (Model 2), in which the three two-way interaction terms (gender X race, gender X age,

race X age) are included in addition to the main effects. Finally, the saturated model is tested (Model 3), in which the three-way interaction (gender X race X age) is added to the previous model (Kennedy, 1983).

The tests for the screening of these four families of models with mode effect groups are presented for the Verbal scale in Table 8. The goodness of fit chi-square statistics provided tests of the amount of discrepancy between the observed cell frequencies and the expected frequencies provided by each model. A large value of chi-square (i.e., one with a probability of less than .05 under a true null hypothesis), indicates that the observed cell frequencies are sufficiently different from those expected that the model is not tenable. The goodness of fit tests in Table 8 indicated that any of the models provide an adequate fit for the data.

In addition to the goodness of fit test, the likelihood ratio chi-square is used to test the difference in fit between the reduced model relative to the model with additional effects included. This test, referred to as a component analysis, is analogous to the more familiar test for a change in R^2 in multiple regression analyses. The component analyses in Table 8 indicated that none of the differences in fit between successive pairs of models proved to be statistically significant. This indicates that the variables, as measured here, are not strongly related to mode effect group.

Table 8

Log-linear Analysis of Mode Effect Group by Gender, Racial/Ethnic Background and Age for Verbal Scale

Model	Goodness of Fit			Component		
	L^2	df	p	L^2	df	p
0	44.50	38	0.22			
1	31.18	26	0.22	13.32	12	0.35
2	2.37	8	0.97	28.81	18	0.06
3	.	0	.	2.37	8	0.97

Table 9 presents the log-linear analysis conducted on the Quantitative scale. The goodness of fit tests presented here indicated that the null model (Model 0) did not provide an acceptable fit to the observed data ($L^2 = 55.44$, $p < .03$), but that any of the other three models fit the data adequately. The component analyses indicated that the difference in fit between the null model and the main effects model (Model 1) is statistically significant, $L^2 = 33.59$, $p < .01$. The addition of two-way interactions (Model 2) to the main effects model did not significantly enhance the degree of fit, nor did the addition of the three-way interaction to the two-way interaction model.

The results of the goodness of fit and component analyses suggested that at least one of the explanatory variables (gender, racial/ethnic background, and age group) was associated with mode effect, but that no significant interactions between these explanatory variables were evident. To determine which explanatory variables were related to mode effect, additional analyses were conducted. Because the three explanatory variables were correlated with each other, these analyses were conducted by testing the fit of models in which only two of these explanatory variables were included. Tests of the difference in the fit of each of these models, relative to the model in which all main effects were included, provide a test of the significance of each explanatory variable after adjusting for the other explanatory variables. As indicated by the component analyses in Table 10, significant effects were obtained for the explanatory variables of gender and age. The variable of examinee race, however, showed no significant relationship with mode effect after adjusting for gender and age.

The bivariate frequency distribution of gender on mode effect for the Quantitative scale indicates that female examinees, relative to males, were more frequently found in the no mode effect group (59%) and the paper mode effect group (83%) compared to the computer mode effect group (56%). Similarly, the bivariate distribution of examinee age and mode effect group for the Quantitative scale indicates that younger examinees, relative to older examinees, were more frequently found in the no mode effect group (82%) and the computer mode effect group (84%), compared to the paper mode effect group (74%).

Table 9

Log-linear Analysis of Mode Effect Group by Gender, Racial/Ethnic Background and Age for Quantitative Scale

Model	Goodness of Fit			Component		
	L^2	df	p	L^2	df	p
0	55.44	38	0.03			
1	21.85	26	0.70	33.59	12	0.01
2	1.33	8	1.00	20.52	18	0.30
3	.	0	.	1.33	8	0.99

Table 10

Component Analysis on L^2-Change for Main Effects of Gender, Racial/Ethnic Background and Age on Mode Effect for Quantitative Scale

Model	Goodness of Fit			Component		
	L^2	df	p	L^2	df	p
All main effects	21.85	26	0.70			
Race & Age only	1.67	28	1.00	20.18	2	0.01
Race & Gender only	15.32	28	0.91	6.53	2	0.05
Age & Gender only	16.74	34	0.54	5.11	8	0.70

The results of the log-linear analysis on the Analytical scale are given in Table 11. The results of the goodness of fit tests indicated that any of the models provide an adequate fit for the data. However, results of the component analyses indicated that the difference in fit between the null model and the main effects model (Model 1) was statistically significant, $L^2 = 25.16$, $p < .05$. Neither the addition of two-way interactions (Model

2) to the main effects model, nor the addition of the three-way interaction (Model 3) to the two-way interaction model significantly enhanced the fit to the observed data. These results suggested that at least one of the explanatory variables (gender, racial/ ethnic background, and age group) was associated with mode effect, but that no significant interactions between these explanatory variables was evident.

Follow-up analyses were performed to determine which explanatory variables were related to mode effect. The component analyses in Table 12 indicate a significant relationship between age and mode effect. The variables of gender and race both failed to show a significant relationship with mode effect after adjusting for the other variables.

Examination of the bivariate distribution of age on mode effect for the Analytical scale indicates that younger examinees, relative to older examinees, were more frequently found in the no mode effect group (84%) and the computer mode effect group (83%), compared to the paper mode effect group (70%).

Table 11

Log-linear Analysis of Mode Effect Group by Gender, Racial/Ethnic Background and Age for Analytical Scale

Model	Goodness of Fit			Component		
	L^2	df	p	L^2	df	p
0	37.13	38	0.51			
1	11.97	26	0.99	25.16	12	0.02
2	1.28	8	1.00	10.69	18	0.90
3	.	0	.	1.28	8	0 99

Table 12

Component Analysis on L^2-Change for Main Effects of Gender, Racial/Ethnic Background and Age on Mode Effect for Analytical Scale

Model	Goodness of Fit L^2	df	p		Component L^2	df	p
All main effects	12.04	26	0.99				
Race & Age only	12.91	28	0.99		.87	2	0.99
Race & Gender only	19.73	28	0.87		7.69	2	0.02
Age & Gender only	26.03	34	0.83		13.99	8	0.09

Computer Experience

The next analysis performed was an investigation into the relationship between mode effect and variety and amount of computer experience. Overall, this sample of examinees was highly computer literate. Only 1% of the examinees in this study had never used a computer before taking the computer test, and only 15% had never used a mouse previously.

Three questions from the computer test survey were used to measure the extent of examinees' computer experience. First, variety of computer experience was measured by examinee response to the question "For what kinds of activities do you use a personal computer?". For this analysis, the variety of types of activities with which an examinee had experience was the point of interest; responses were therefore dichotomized as single type versus multiple types of activities. Responses to this question confirmed the high level of computer experience in this sample. Over 93% of examinees indicated that they used computers for wordprocessing and fully 72% had experience with two or more types of software packages.

Frequency of computer use was measured by the question "How often do you use a personal computer?"; frequency of mouse use was measured by the question "How often have you used a mouse on a

personal computer?". Responses to both these questions were on a 4-point scale ranging from "1=Never" to "4=Routine Use."

For the Method 1 approach, data were subjected to a three-factor MANOVA; no significant differences in scores were found between levels of the computer experience variables examined (see Table 13).

Table 13

MANOVA of Residual Difference Scores on Verbal, Quantitative, and Analytical Scales by Variety of Computer Experience, Frequency of Computer Use, and Frequency of Mouse Use

Source	df	Wilks' Lambda	F
Variety of Computer Experience (V)	3, 1078	1.00	.37
Frequency of Computer Use (A)	9, 2624	.99	.85
Frequency of Mouse Use (M)	9, 2624	.99	1.27
V * A	6, 2156	1.00	.43
V * M	9, 2624	1.00	.19
A * M	18, 3050	.99	.84
V * A * M	12, 2852	.99	.81

Under Method 2, data were subjected to individual chi-square tests (Table 14). Only one significant result was found. On the Analytical scale there was an association between mode effect and frequency of mouse use, although the strength of the relationship was modest.

An examination of the related two-way frequency distribution for the Analytical scale included in Table 15 suggests that the source of the significant chi-square can be found in dependence between mode effect grouping and the "never" response to the survey question regarding frequency of mouse use. Those examinees who had never used a mouse prior to the examination were more likely to fall into the paper mode effect group, while those with regular or routine mouse use were more likely to fall into the computer mode effect group.

Table 14

Chi-Square Tests of Variety and Frequency of Computer Use by Mode Effect Group

Test	df	Chi-Square	Cramer's V
Variety of Computer Experience			
Verbal	2	3.38	.06
Quantitative	2	0.89	.03
Analytical	2	4.57	.06
Frequency of Computer Use			
Verbal	6	5.59	.05
Quantitative	6	5.05	.05
Analytical	6	7.66	.06
Frequency of Mouse Use			
Verbal	6	7.60	.06
Quantitative	6	11.04	.07
Analytical	6	19.90*	.10

* $p < .01$.

Table 15

Two-Way Frequency Distribution of Frequency of Mouse Use on Mode Effect Group by Scale

Mode Effect	Frequency of Mouse Use				
	Never	Rarely	Regular	Routine	No Response
Verbal Scale					
Computer	25 (16)	46 (29)	53 (33)	34 (21)	1 (1)
No Effect	114 (14)	289 (37)	229 (29)	151 (19)	5 (1)
Paper	31 (19)	53 (32)	42 (25)	39 (23)	2 (1)
Quantitative Scale					
Computer	14 (13)	32 (30)	32 (30)	26 (25)	1 (1)
No Effect	139 (15)	310 (34)	264 (29)	188 (21)	6 (1)
Paper	17 (17)	46 (45)	28 (27)	10 (10)	1 (1)
Analytical Scale					
Computer	7 (7)	29 (31)	40 (43)	16 (17)	2 (2)
No Effect	137 (15)	325 (36)	254 (28)	193 (21)	6 (1)
Paper	26 (25)	34 (32)	30 (29)	15 (14)	0 (0)

Note. Percentages are given in parentheses. N = 1114.

Test Mode Preference

The next analysis considered an additional computer use variable: test mode preference. Test mode preference was measured through responses to the question "If there were a computer test and a paper-and-pencil test with the same questions, which would you prefer to take?". Overall, 61% of the respondents indicated that they would prefer the computer mode, while 29% indicated the paper mode, and 11% reported no preference.

The relationship between mode effect and examinees' test mode preference was analyzed under the Method 1 approach using a one-way MANOVA of examinees' residualized difference scores on the three examination scales by test mode preference. No significant main effect was found for test mode preference, $F(6, 2150) = 1.80$, $p > .05$, Wilks' Lambda = .99.

Data were analzyed using the Method 2 approach through individual chi-square analyses. Significant test mode preference effects were found for both the Verbal and Analytical scales of the examination, with $\chi^2(4, N = 1082) = 10.95$, $p < .05$ and; $\chi^2(4, N = 1082) = 14.94$, $p < .01$, respectively. No significant effect was found for the Quantitative scale, $\chi^2(4, N = 1082) = 6.06$, $p > .05$.

Visual examination of the two-way frequency distributions of mode effect groups by test mode preference (Table 16) suggested that on both the Verbal and Analytical scales examinees tended to prefer the test mode in which they performed better.

Table 16

Two-Way Frequency Distribution of Test Mode Preference and Mode Effect on Verbal, Quantitative, and Analytical Residual Difference Scores

Mode Effect	Test Mode Preference: Computer	No Preference	Paper	No Response
Verbal Scale				
Computer	112 (70)	13 (8)	33 (21)	1 (1)
No Effect	441 (56)	89 (11)	233 (30)	25 (3)
Paper	103 (61)	61 (8)	45 (27)	6 (4)
Quantitative Scale				
Computer	68 (65)	9 (8)	20 (19)	8 (8)
No Effect	522 (58)	97 (11)	266 (29)	22 (2)
Paper	66 (65)	9 (9)	25 (24)	2 (2)
Analytical Scale				
Computer	68 (72)	77 (8)	14 (15)	5 (5)
No Effect	523 (57)	103 (11)	263 (29)	26 (3)
Paper	65 (62)	5 (5)	34 (32)	1 (1)

Note. Percentages are given in parentheses. N = 1114.

Test Strategy Preference

The next analysis considered the relationship between mode effect and test strategy preference. Test strategy preference was measured as examinee reaction to testing constraints on the final section of the computer version of the examination. Examinees were asked the survey question: "You were not permitted to use the 'Review' option during the last (seventh) section. What was your reaction to this testing rule?". Only 20% of the examinees who responded to this question indicated "did not care", while 41% indicated that they found these restrictions "somewhat frustrating" and 39% found the constraints "very frustrating".

The Method 1 procedure for investigating this question utilized a one-way MANOVA of residualized difference scores on the examination scales by test strategy preference. A significant main effect for test strategy preference was found, $F(6, 2040) = 3.01$, $p < .01$, Wilks' Lambda = .98. Follow-up one-way ANOVAs revealed significant effects for the Verbal and Analytical scales, $F(2, 1023) = 3.44$, $p < .05$, and $F(2, 1023) = 5.55$, $p < .01$, respectively. No significant effect was found for the Quantitative scale, $F(2, 1023) = 2.30$, $p > .05$. For both the Verbal and Analytical scales, Tukey's post hoc test revealed significant pairwise differences ($p < .05$) between examinees who responded "did not care" and examinees in the two other groups. The lower mean residual difference score for examinees in this group suggests that they performed less well on the computer version than their paper version scores would have predicted, as compared to examinees in the two "frustrated" groups. No significant differences were found between examinees in the two groups "somewhat frustrated" and "very frustrated" on either the Verbal or Analytical scales.

Individual chi-square tests were employed for the Method 2 approach. A significant result was found for the Analytical scale, $\chi^2(4, N = 1027) = 14.76$, $p < .01$. An examination of the two-way frequency distribution associated with this test (see Table 17) revealed that twice as many examinees who responded "did not care" came from the paper mode group as from the computer mode group. No significant results were found for the Verbal or Quantitative scales, $\chi^2(4, N = 1027) = 3.52$, $p > .05$ and $\chi^2(4, N = 1027) = 7.86$, $p > .05$ respectively.

Table 17

Two-Way Frequency Distribution of Test Strategy Preference and Mode Effect on Residual Difference Scores

Mode Effect	Test Strategy Preference: Did Not Care	Somewhat Frustrating	Very Frustrating	No Response
Verbal Scale				
Computer	24 (15)	67 (42)	59 (37)	9 (6)
No Effect	148 (19)	295 (37)	282 (36)	63 (8)
Paper	37 (22)	61 (37)	54 (32)	15 (9)
Quantitative Scale				
Computer	16 (15)	45 (43)	36 (34)	8 (8)
No Effect	164 (18)	341 (38)	329 (36)	73 (8)
Paper	29 (29)	37 (36)	30 (29)	6 (6)
Analytical Scale				
Computer	15 (16)	38 (41)	37 (39)	4 (4)
No Effect	160 (17)	354 (39)	326 (36)	75 (8)
Paper	34 (32)	31 (30)	32 (30)	8 (8)

Note. Percentages are given in parentheses. N = 1114.

Test Flexibility

The final area of analysis in this study was the relationship between examinees' test taking strategy (operationally defined as the tendency to omit or review items) and flexibility. Data for this question were obtained entirely from the computer Verbal sections of the examination, since flexibility was not manipulated elsewhere. In the first six sections of the examination subjects were able to omit and revise items. The seventh section of the test was non-flexible, or constrained; examinees could not omit items, and they were not permitted to review or revise an item once that item was completed. This non-flexible section was a Verbal section and was otherwise completely parallel to the other two Verbal sections. For the non-flexible condition, the dependent variable was the simple number correct score. For the flexible condition, the mean number correct score of the two flexible Verbal sections was the dependent variable. Means and standard deviations for these scores are given by mode effect group and flexibility in Table 18.

Table 18

Means and Standard Deviations of Verbal Number Correct Score by Mode Effect Group and Level of Flexibility

Mode Effect	Level of Flexibility	
	Flexible	Non-Flexible
Computer Mode Effect		
M	28.26	30.81
SD	4.76	4.73
No Mode Effect		
M	25.83	29.65
SD	5.78	5.91
Paper Mode Effect		
M	19.92	26.45
SD	4.88	5.56

Note. N = 1114.

Data were analyzed through two repeated measures ANOVAs. First was a one-between, one-within subjects ANOVA with mode effect group as the between-subjects factor and level of flexibility as the within-subjects (repeated measures) factor. A significant main effect for flexibility was found, as was a significant interaction between flexibility and mode effect (Table 19).

Table 19

One-Between, One-Within Subjects ANOVA of Verbal Number Correct Scores by Mode Effect Group and Flexibility

Source	df	MS	F
Between Ss			
Mode Effect (M)	2	3747.37	65.26*
Error (S/M)	1111	57.42	
Within Ss			
Flexibility (F)	1	6281.44	1135.99*
MF	2	356.89	64.54*
Error (SF/M)	1111	5.53	
Total	2227		

* $p < .01$.

As a follow-up to the interaction effects, correlated means t-tests were computed on the difference between the non-flexible and the flexible conditions for each level of mode effect group. Contrary to anticipated results, the mean examinee performance was significantly higher under the non-flexible condition than under the flexible condition for all three levels of mode effect: for the computer mode effect group, $t(159) = 10.99$, $p < .01$; for the no mode effect group, $t(788) = 33.19$, $p < .01$; and for the paper mode effect group, $t(167) = 22.31$, $p < .01$.

In order to determine whether those examinees whose test taking strategy includes more omissions and revisions would be more greatly affected by constrained conditions, a three-between, one-within subjects ANOVA

was conducted. The between-subjects factors of this ANOVA, in addition to mode effect, included categorical measures of item omits and item reviews; level of flexibility was again the within-subjects factor.

The measure of item reviews was based on responses to the question: "For how many questions did you use the 'Review' screen?" Response options to this question were on a 6-point scale, ranging from "1=None" to "6=Almost all." These six categories were collapsed to three (None, A few - 1/4, and 1/2 to Almost All). The measure of item omits was obtained by dichotomizing the actual count of Verbal items omitted by examinees. Examinees were then categorized as Low (0 - 2) or High (3 or more) in omitting.

The addition of these measures did not add a great deal to the analysis. The behaviors as measured here did not appear to be related to mode effect, as evidenced by the lack of interaction between mode effect and either item omits or item reviews (see Table 20). The only significant two-way interaction, flexibility by omits, revealed a significant improvement in scores for examinees who were high in the tendency to omit, when the opportunity to omit items was not available.

Table 20

Three-Between, One-Within Subjects ANOVA of Verbal Number Correct Score on Mode Effect Group, Omitting, Reviewing, and Flexibility

Source	df	MS	F
Between Ss			
Mode Effect (M)	2	266.03	4.77**
Omits (O)	1	239.20	4.29*
Reviews (R)	2	471.96	8.46**
M * O	2	60.92	1.09
M * R	4	34.53	0.62
O * R	1	2.14	0.04
M * O * R	2	9.57	0.17
Error S/MOR	1031	55.81	
Within Ss			
Flexibility (F)	1	364.67	66.49**
F * M	2	36.62	6.68**
F * O	1	22.90	4.18*
F * M * O	2	2.87	0.59
F * R	2	11.75	0.12
F * M * R	4	6.27	0.33
F * O * R	1	17.78	0.07
F * M * O * R	2	8.99	0.19
Error (F/MOR)	1031	5.48	
Total	2091		

* $p < .05$. ** $p < .01$.

Discussion

The data in this study demonstrated mode effect and supported the conception of a small subset of examinees whose performance was more affected by test mode than was that of the total sample of examinees (see, for example, Table 2). As always, the results should be interpreted in light of limitations on the research design. Limitations in this research include the confounding of test mode with order: all examinees took the paper version of the examination first and the computer version last. Flexibility also is confounded with order. Additionally, certain of the variables were measured less precisely than would be optimal. As an example of better measures of variables, the test taking strategy of item reviews could be more appropriately obtained in future research through test software in which an actual count is taken of the number of times an examinee views each item.

The search for examinee characteristics that could explain the occurrence of mode effect yielded inconsistent results. The variables investigated in this study showed only relatively weak relationships with mode effect. Results under the Method 1 and the Method 2 approaches were largely parallel, as the summary of results in Table 21 shows. It is possible that more appropriate variables exist, or that measurement of the variables investigated in this study could be improved; either of these cases could yield evidence of stronger relationships between examinee characteristics and mode effect.

Demographic Variables

An overall MANOVA on the demographic variables of gender, racial/ethnic background, and age showed significant main effects for gender and race, and a significant race by age interaction (Method 1). However, these results varied across the three scales of the GRE General Test and some of them failed to hold up under pairwise, post hoc analyses.

Significant effects were found for gender on both the Verbal and Quantitative scales. Interestingly, on the Verbal scale, the gender by race interaction effect showed the mean residual score for non-White males to be negative, while for non-White females it was positive. This indicates that, based on their performance on the paper version of the Verbal scale, males on average performed less well on the computer version of that scale than would be expected. On the Quantitative scale the pattern of performance was reversed; a main effect for gender was found in which males had higher mean residual difference scores than females. That is, the performance of males on the

Table 21

Summary of Results for Statistical Tests of Significance Under Method 1 and Method 2 by Research Question

Statistical Results			
Method 1		Method 2	
Demographics			
MANOVA	sig	Log-linear	
ANOVA - Verbal	sig	Verbal	ns
Quantitative	sig	Quantitative	sig
Analytical	sig	Analytical	sig
Computer Experience			
MANOVA	ns	Chi-Square	
Verbal		Verbal	ns
Quantitative		Quantitative	ns
Analytical		Analytical	sig
Test Mode Preference			
MANOVA	ns	Chi-Square	
Verbal		Verbal	sig
Quantitative		Quantitative	ns
Analytical		Analytical	sig
Test Strategy Preference			
MANOVA	sig		
ANOVA		Chi-Square	
Verbal	sig	Verbal	ns
Quantitative	ns	Quantitative	ns
Analytical	sig	Analytical	sig
Tendency to Omit/Review (Verbal)			
Main Effect for Flexibility	sig	Flexibility x Mode Effect	sig
Flexibility x Omits	sig	Flexibility x Omits x Mode Effect	ns
Flexibility x Reviews	ns	Flexibility x Reviews x Mode Effect	ns

computer version of the Quantitative scale was better than their paper Quantitative scores would predict, while females performed less well on the computer Quantitative scale than their paper Quantitative scores would predict.

On both the Verbal and Quantitative scales, a race by age interaction revealed a significant difference between the residual difference scores of older and younger Black examinees. Older Black examinees performed less well on the computer version than their paper version scores predicted, as compared to younger Black examinees. On the Analytical scale, a main effect for race, followed by pairwise comparisons between races, revealed a significant difference between Asians and Blacks. The difference between residual scores for Asians and Blacks suggested that, based on their paper Quantitative scores, Asians did better on the computer Quantitative scale as compared to Blacks.

Results of the log-linear analysis (Method 2) revealed main effects for both gender and age in the Quantitative scale and a main effect for age in the Analytical scale. Follow-up examinations of bivariate frequency distributions for the Quantitative scale suggested that females were found in the paper mode effect group and no mode effect group to a greater extent than in the computer mode effect group. Younger examinees were found in greater frequencies in the no mode effect group and the computer mode effect group, relative to the paper mode effect group for both the Quantitative and Analytical scales.

Interpretation of the mixed results found for demographic variables is difficult. It may be that mode effect is not directly related to any of the demographic variables investigated. Perhaps, instead, mode effect is an indirect outcome of some other variable, such as test anxiety, and its interaction with certain demographic attributes. Such an interaction might explain the differing patterns of performance by males and females on the Verbal and Quantitative scales.

Computer Use Variables

Minimal results were found in this study to support the relationship between computer experience and mode effect. When the data were subjected to a three-factor MANOVA (the Method 1 approach), no significant differences in residual difference scores were found between levels of the computer experience variables examined. Only one significant result was found for the Method 2 approach of individual chi-square tests: on the Analytical scale there was an association between mode effect and frequency of mouse use. Examination of the two-way

frequency distribution of these variables suggested that examinees with no prior mouse use were over-represented in the paper mode effect group, while those with regular or routine mouse use were more strongly represented in the computer mode effect group.

Similarly, few results were found for the relationship between test mode preference and mode effect. A MANOVA of examinees' residualized difference scores on the three examination scales by test mode preference, the Method 1 approach, revealed no significant effect. The individual chi-square analyses conducted as the Method 2 approach revealed significant test mode preference effects for both the Verbal and Analytical scales. For both scales, the relationship appeared to be a tendency for examinees to prefer the test mode in which they performed better.

Perhaps it was because of the high level of computer experience in this sample that investigation of computer use variables failed to show any strong results. The high representation of computer experience in this sample was somewhat expected, given that participation in the computer test pilot study was entirely voluntary. Nevertheless, the lack of subjects with no computer experience makes analysis of the relationship between mode effect and computer use variables very difficult.

Lee (1986) found that while examinees with no computer experience were negatively affected by computer administration of a test, those examinees categorized as "low" on computer experience were not. It may be that as Wainer, Dorans, Green, Mislevy, Steinberg, and Thissen (1990) have suggested, computer tests require so few computer skills that frequency and variety of computer experience do not impact mode effect. On the other hand, the extent of computer experience which is beneficial to an examinee may vary across different computer test conditions; the impact of computer experience on mode effect may depend upon the complexity of the test administration software. Test administration software with a cumbersome interface has been noted for increased testing time and increased examinee anxiety (Harvey, 1987; Johnson & Johnson, 1981). Differences in testing software could confound the relationship of computer experience to mode effect.

Test Taking Strategy Preferences

Examinee test taking strategy preferences was considered to be an important variable for analysis in this study because the two test modes of computer and paper versions may match up differentially with different strategy

approaches. If examinees strongly prefer one set of strategies, and the test mode processes conform well to those strategies, then those examinees may be aided in their overall performance under that test mode. Conversely, if a given test mode restricts the use of those strategies but provides a good match for other strategies, then those examinees who are able to adjust to these processes and apply new strategies will be better off than those examinees who are not able to do so.

This study was specifically concerned with test taking strategy as it applies to differences in levels of flexibility across test modes. When flexibility is discussed as an issue in computer testing it is often from the perspective that not allowing examinees to omit and revise items may penalize them, compared to their potential performance on the same test in paper form (Spray et al., 1989; Ward et al., 1989). In addition, the literature on answer changing on paper tests consistently shows that examinees are likely to improve their scores when they change their responses to test items (Green, 1981; Matter, 1986; Mueller & Shwedel, 1975; Penfield & Mercer, 1980). This would tend to support the need for providing flexibility in computer tests. The results of this study, however, would at first glance appear to support the opposite position; the overall mean score for the non-flexible Verbal section of the examination was higher than for the flexible sections. Contrary to results found in other studies on flexibility (Spray et al., 1989; Ward et al., 1989), performance was significantly higher in the non-flexible condition; although this was true at all three levels of mode effect, the difference was especially marked in the paper mode effect group.

Despite the overall performance increase in the non-flexible section, the non-flexible condition was a source of frustration to most examinees. A total of 80% of the examinees indicated that they found the constraints on omitting and revising to be frustrating. The relationship between test strategy preferences and mode effect was examined through Method 1 ANOVAs and significant effects were found for the Verbal and Analytical scales. Those examinees who indicated "did not care" had a significantly lower mean residual score, indicating that their computer test performance was not as high as their paper test performance predicted, as compared to examinees who selected "somewhat frustrated" or "very frustrated". Under Method 2, a significant chi-square test was also found for the Analytical scale. Examination of the bivariate frequency distribution between test mode preference and mode effect groups also suggested that the "did not care" response was associated with the paper mode effect group. One

possible interpretation of these results is that those examinees most comfortable with the sequential pattern of responding to items forced by non-flexible conditions would also be those whose performance would be higher in typical paper administrations.

Test taking strategy and test flexibility may be related in more complex ways than have yet been fully studied. Issues concerning speededness, the accessibility of the test administration software, and the structure and organization of the examination may all contribute to delineating the best strategy for a given test administration. More precise measures of examinees' item omitting and reviewing behaviors ought to be collected to further investigate the relationship of these strategy preferences with mode effect under the constrained conditions often found in computer testing.

Summary

The presence of a mode effect was detected in the data used in this study. Specifically, while the majority of examinees were not affected by test mode, one subset of examinees performed better on the computer version, and a second subset of examinees performed better on the paper version of the test. However, the search for examinee characteristics that explain the occurrence of mode effect yielded mixed results. The variables investigated in this study showed only relatively weak relationships to mode effect. Further investigation tailored to this question should be conducted in order to determine those variables which distinguish those examinees whose performance is affected by mode of test administration from those whose performance is not.

References

Becker, H. J., & Sterling, C. W. (1987). Equity in school computer use: National data and neglected considerations. Journal of Educational Computing Research, 3, 289-311.

Bugbee, A. C., & Bernt, F. M. (1990). Testing by computer: Findings in six years of use. Journal of Research on Computing in Education, 23, 87-100.

Buhr, D. C., & Legg, S. M. (1989). Development of an adaptive test version of the College Level Academic Skills Test. (Institute for Student Assessment and Evaluation, Contract No. 88012704). Gainesville, FL: University of Florida.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), Educational Measurement (3rd ed, pp. 367-408). New York: Macmillan.

Committee to Develop Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Eaves, R. C., & Smith, E. (1986). The effect of media and amount of microcomputer experience on examination scores. Journal of Experimental Education, 55, 23-26.

Educational Testing Service (1992, April). Introduction and overview of the GRE Computer-Based Testing Project. In Clyde Reese (Organizer-Moderator), Development of a Computer-Based Test for the GRE General Test. Symposium conducted at the meeting of the National Council on Measurement in Education, San Francisco.

Green, B. F. (1991). Guidelines for computer testing. In T. B. Gutkin & S. L. Wise (Eds.), The Computer and the Decision-Making Process (pp 245-254). Hillsdale, NJ: Lawrence Erlbaum.

Green, K. (1981). Item-response changes on multiple-choice tests as a function of test anxiety. Journal of Experimental Education, 49, 225-228.

Harvey, A. L. (1987). Differences in response behavior for high and low scorers as a function of item presentation on a computer assisted test. Unpublished doctoral dissertation, University of Nebraska, Lincoln.

Johnson, D. F., & Mihal, W. L. (1973). Performance of blacks and whites in computerized versus manual testing environments. American Psychologist, 28, 694-699.

Johnson, D. F., & White, C. B. (1980). Effects of training of computerized test performance in the elderly. Journal of Applied Psychology, 65, 357-358.

Johnson, J. H., & Johnson, K. N. (1981). Psychological considerations related to the development of computerized testing stations. Behavior Research Methods & Instrumentation, 13, 421-424.

Kennedy, J. J. (1983). Analyzing Qualitative Data: Introductory Log-Linear Analysis for Behavioral Research. New York: Praeger Publishers.

Koch, B. R., & Patience, W. M. (1978). Student attitudes toward tailored testing. In D. J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology.

Lee, J. A. (1986). The effects of past computer experience on computerized aptitude test performance. Educational and Psychological Measurement, 46, 721-733.

Llabre, M. M., & Froman, T. W. (1987). Allocation of time to test items: A study of ethnic differences. Journal of Experimental Education, 55, 137-140.

Matter, M. K. (1986, April). Eenie, meenie, minie, mo -- change this answer -- yes or no? Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Mazzeo, J., & Harvey, A. L. (1988). The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature (College Board Rep. No. 88-8, ETS RR No. 88-21). Princeton, NJ: Educational Testing Service.

McPhail, I. P. (1985). Computer inequities in school uses of microcomputers: Policy implications. Journal of Negro Education, 54, 3-13.

Moe, K. C., & Johnson, M. F. (1988). Participants' reactions to computerized testing. Journal of Educational Computing Research, 4, 79-86.

Mueller, D. J. & Shwedel, A. (1975). Some correlates of net gain resultant from answer changing on objective achievement test items. Journal of Educational Measurement, 12, 251-254.

Penfield, D. A. & Mercer, M. (1980). Answer changing and statistics. Educational Research Quarterly, 5, 50-57.

Rocklin, T., & O'Donnell, A. M. (1987). Self-adapted testing: A performance-improving variant of computerized adaptive testing. Journal of Educational Psychology, 79, 315-319.

Sachar, J. D., & Fletcher, J. D. (1978). Administering paper-and-pencil tests by computer, or the medium is not always the message. In D. J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference. Wayzata, MN: University of Minnesota.

Sorensen, H. B. (1985). Cognitive ability tests. AEDS Monitor, 24, 22-26.

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. Journal of Educational Measurement, 26, 261-271.

Wainer, H., Dorans, N. J., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). Future challenges. In H. Wainer (Ed.), Computerized Adaptive Testing: A Primer (pp 233-272). Hillsdale, NJ: Lawrence Erlbaum.

Ward, T. J., Jr., Hooper, S. R., & Hannafin, K. M. (1989). The effects of computerized tests on the performance and attitudes of college students. Journal of Educational Computing Research, 5, 327-333.

Wise, S. L., Barnes, L. B., Harvey, A. L., & Plake, B. S. (1989). Effects of computer anxiety and computer experience on the computer-based achievement test performance of college students. Applied Measurement in Education, 2, 235-241.

Wise, S. L., & Plake, B. S. (1989). Research on the effects of administering tests via computers. Educational Measurement: Issues and Practice, 3, 5-10.